Results and lessons learned from the quantitative evaluation of road detection and tracking algorithms

D. Dufourd and A. Dalgalarrondo
DGA/Centre Technique d'Arcueil
16bis, av. Prieur de la Côte d'Or, 94114 Arcueil Cedex, France,
Delphine.Dufourd@etca.fr, Andre.Dalgalarrondo@etca.fr

ABSTRACT

In a previous presentation at PerMIS'02, we described a methodology to assess image processing algorithms dedicated to unstructured road detection and tracking. In this paper, we present our first application of this methodology on six algorithms using a database containing about 20,000 images. The main scope of this article consists in presenting the results and the lessons learned from this evaluation.

KEYWORDS: *road detection and tracking, performance evaluation, image processing algorithm assessment, ground robotics, outdoor navigation.*

1 INTRODUCTION

Since the beginning of the 1990's, the French defense procurement agency (Délégation Générale pour l'Armement) has launched several advanced studies dedicated to ground robotics. Among them, the prospective program called PEA Robotique, which started in December 1999, aims at developing a generic teleoperation kit as well as autonomous functions for military unmanned vehicle navigation, such as autonomous road following, beacon or vehicle tracking and scene analysis. In this context, the Centre Technique d'Arcueil (CTA) of the Délégation Générale pour l'Armement (DGA) is currently conducting an evaluation of six existing image processing algorithms for unstructured road detection and tracking. The goal of this evaluation is to compare different road detection and tracking algorithms in a reproducible and quantitative way so as to direct future developments in autonomous outdoor navigation. It should allow us to determine the most promising techniques and possibly find orthogonal strengths between the algorithms so as to conceive hybrid and potentially more efficient methods. Moreover, it should help us to quantify the performances that need to be reached by future algorithms.

The outcome of the evaluation is described in the following sections. Section 2 briefly recalls existing assessment methodologies for image processing algorithms, as well as previous work on the evaluation of road tracking algorithms. Section 3 describes our assessment methodology and section 4 presents some results of the evaluation for the road tracking algorithms. Section 5 explains how the assessment results can be used for further developments and section 6 focuses on lessons learned about the whole evaluation process. Finally, section 7 concludes and outlines future developments.

2 PREVIOUS WORK ON EVALUATION

In the last years, the image processing community has started to develop evaluation methods in order to be able to compare quantitatively the huge number of algorithms available after these last decades of research. Such an approach is very important for those who use image processing as a part of their research, like roboticists, since it provides a guide based on performance among the overwhelming number of available algorithms. However, it should be noted that such an approach is quite new. Up to very recently, algorithms were not evaluated quantitatively, but only qualitatively on various criteria such as the neatness of their design or the sophistication of the underlying mathematical theoretical tools. Most experiments are conducted by human experts and lack any automation. The performance of the algorithms then depends on the know-how and the personal experience of the expert. Fortunately, the situation is changing and there are always more special issues in journals or conferences focusing on image processing assessment issues.

Although a wide variety of vision-based road following algorithms have been proposed and implemented over the last two decades, few techniques have been developed to assess their quality. Far too many articles rely on qualitative results, exhibiting a handful of example images to illustrate the performance of the algorithms while real applications would mean processing millions of image without making any serious error [9]. In many cases, the efficiency of road following algorithms is only characterized by the speed achieved by the whole autonomous system or the time elapsed between two manual interventions [1]. However, using such global characterizations, it seems difficult to determine exactly what makes the system efficient and

Figure 1: Examples of road images of the DGA testing facilities near Angers.

what could be improved to make it better. Algorithms performing 3-D road reconstruction have been evaluated in different ways: indirect numerical tests comparing real and estimated road width and vehicle speed [6], task-oriented metrics using ground truth on both synthetic and real data [2], etc. However, manual 3-D reconstruction appears too time-consuming if the evaluation is to be performed on numerous data. A few research studies focus on automating the measurement of ground truth for the evaluation of vision-based lane sensing : development of a specific device (a side-looking camera and a separate vision system) to measure the offset between the vehicle and the lane [7], simulation of various precipitation rates on real data using a detailed calibration of the imaging system [5], etc. Finally, a few studies select only a well-defined aspect of the system performance in a single class of lane-sensing techniques to enable the automatic extraction of the ground truth in a well-defined and simple context [9] (see [3] for a more detailed description of assessment methodologies of road tracking algorithms).

3 ASSESSMENT METHODOLOGY AND TOOLS

Given the variety of algorithms to be tested, the assessment methodology has to be flexible and generic enough. We have opted for an evaluation based on a ground truth, which implies the use of an image sequence database associated with the corresponding ground truth. Moreover, since we aim at a quantitative evaluation, we had to develop several metrics in order to measure and compare the performances of the different algorithms.

The database includes both the images that compose the input of the image processing algorithms and the ground truth suited to the final task to assess. For our purpose, we needed images of ill-structured roads and trails taken from a vehicle whose size and mobility are close to the targeted UGV. Collecting these images is quite easy with nowadays technologies. The two main difficulties are related to the representativity of the images with respect to the missions and the environment of the UGV, and the constitution of the ground truth. To address the first issue, we specified two kinds of scenarios: six general ones presenting an increasing difficulty level for road extraction and twelve special scenarios which are dedicated to road particularities. In the first case, one obtains homogeneous sequences of images in order to assess an algorithm along a sequence with a low risk of an irreparable failure on some images. Two categories of roads have been defined, each corresponding to three scenarios: tarmac roads and gravel-mud roads. For each category, three levels of difficulty have been determined. In the second case, the special scenarios make it possible to evaluate the algorithm in harsh conditions (hairpin bends on different kinds of soil, abrupt road widening, puddles, slough, road sides with parked vehicles, changing soil, transversal and longitudinal "disturbing" road markings, sequences where the vehicle enters or leaves the road, etc.). Moreover, each general scenario was recorded under three different illumination and weather conditions. This process lead to a first version of the database containing around 20,000 images. Besides, to deal with the constitution of the ground truth, we wrote a detailed specification which guides the human operator in charge of drawing the road boundaries in each image. In order to facilitate this long and dull job, we also created a program with a dedicated interface which speeds up the ground truth definition process. Broadly speaking, it manages the name and numbering conventions of the images and ground truth files of a sequence and allows, on a new image, an easy modification of the ground truth defined on the previous image.

Numerous authors underlined the need for multiple metrics in image processing algorithms assessment, so that users can consider different aspects of these algorithms and choose the one which is best suited to their application [8]. Following this point of view, we propose nine different metrics, computed in the 2-D image space, which aim at assessing geometric accuracy as well as a good global correspondence between the ground truth and the output of the algorithms. Among the various metrics available, we can distinguish contour-oriented and region-oriented metrics, which reflect the dual approaches to image segmentation (see Fig. 2 for the notations). Before computing most contour-oriented metrics, we need to perform a matching procedure between the reference road edge and the result of the algorithm. We chose the so-called "buffer method" in

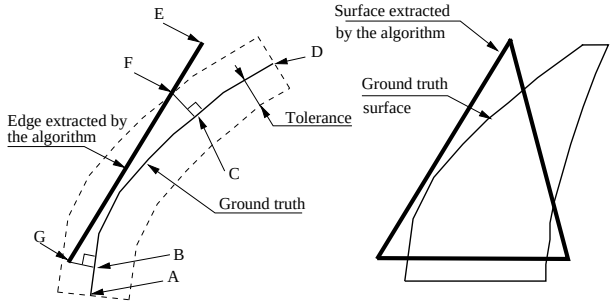

Figure 2: Notations for metrics.

which every portion of the boundary lying within a certain distance (i.e. the width of the buffer) from the reference boundary is considered as matched.

The two first contour-oriented metrics count up the number of pixels which have been misclassified: the completeness metric ($m_1 = BC/AD$) computes the difference between the length of the result judged as valid (within the tolerance buffer) and the length of the ground truth, while the correction metric ($m_2 = GF/GE$) determines what portion of a result lies within the tolerance area. We also define a combination of the previous metrics: $m_3 = m_1 \times m_2$. m_1, m_2 and m_3 equal 1 when the result is perfect for each metric. The fourth contour-oriented metric $m_4 = \Sigma_G^E dist(algorithm, ground\,truth)/length(GE)$ computes the average distance between the reference and result edges. m_4 equals 0 when the result is perfect for this metric. We have added a signed distance metric ($m_{4s} = \Sigma_G^E signed\text{-}horizontal\text{-}distance(algorithm, ground \quad truth)$) which is not symmetrical with respect to the road edges. m_{4s} makes it possible to determine whether the algorithm is rather optimistic (the edges tend to be detected outside the road surface) or pessimistic (the edges are mostly detected inside the road surface). Using both the right and left edges, it also enables to compare the result and reference central road axes. Contour-oriented metrics provide detailed information about the geometric accuracy of the algorithms. However, in sharp turns or on very irregular paths, pessimistic algorithms using a very simple model (a triangle for instance) risk being severely penalized by these metrics even if they find a driveable area within the boundaries of the real road. As a result, we have defined two other metrics based on surface, which measure the frequency of incorrect classification of pixels in the image. The first one is related to completeness ($m_5 = |S_{result} \bigcap S_{ground\,truth}|/|S_{ground\,truth}|$) and the second one to correctness ($m_6 = |S_{result} \bigcap S_{ground\,truth}|/|S_{result}|$). Finally, we define two additional combinations of metrics: $m_7 = m_5 \times m_6$ and $m_8 = m_3 \times m_7$. m_5, m_6, m_7 and m_8 equal 1 when the result is perfect for each metric.

The computation of the metrics given the algorithms and the database is performed automatically using the SENA platform which was developed by Cril Ingenierie under CTA specification and supervision. SENA is a customized software environment for fast algorithm implementation and evaluation of a wide range of applications. It helps in assembling image processing operators and replaying the experiments on large amounts of images. It makes it possible to incorporate tools into the operator sequence in order to measure or visualize partial results. Thus, SENA is able to organize and execute sequences of operators of different types (source code, shell scripts, binaries, libraries) and origins (operators which were developed specifically for the platform or not). The only constraint is that all the operators must be executed on the same host computer. Practically, SENA runs on an SMP computer (Sun Enterprise 10,000 with 32 processors) to cope with the huge amounts of data and range variations of the algorithm parameters (see [3] for a more detailed description of the whole evaluation methodology).

4 SOME RESULTS

A public consultation enabled us to acquire six existing algorithms developed either by French laboratories (LASMEA, LCPC, MINES, CTA) or by French companies (PG:ES). To perform their road-tracking task, these algorithms rely on various strategies (see table 1 for a global description and [4] for more details). Some of them resort to edge detection, others to region classification, and one of them combines both techniques. They operate either on gray-level or on color images. The 2-D geometric models of the road range from triangles or simple straight lines to parabolas and polygonal edges.

Some algorithms depend on different parameters: for instance, the first algorithm from the LASMEA relies on a Mahalanobis distance threshold. To tune these parameters, we have used a single surface metric which computes the percentage of pixels which have been misclassified : $m_9 = 1 - 2|S^r_{result} \bigcap S^r_{ground\,truth}|/|S^r_{result} + S^r_{ground\,truth}|$ where S^r is the surface restrited between 0.6 and 0.8 heigth of the image. This measure concentrate on the most important part of the image that is not too close to the vehicle

	Road model	Road extraction strategy
LASMEA 1	2 straight line segments ; road width, lateral position of the vehicle	Supervised gray-level classification of the image using a road prototype extracted from the previous image. Median least square technique to determine the road edges. Kalman filter to estimate 4 position parameters.
LASMEA 2	Polygonal edges (9 vertices on each side at fixed heights)	Kalman filtering to estimate the abscises of each vertex. Road segments are extracted by a median least square technique. At each image, a hypothesis tree enables to select successively the best observations to update the road estimate.
LCPC	2 straight line segments (end points at fixed heights)	Unsupervised monodimensional classification based on chromatic saturation and thresholding. Extraction of the road edges by a median least square method.
MINES Fontainebleau	2 straight line segments	Mathematical morphology : watershed segmentation on a color gradient image, with different levels of hierarchy. Exploitation of temporal consistence throughout the sequence.
PG:ES	Triangular model	Variant of the classical SCARF algorithm from CMU. Supervised color image classification with two road classes ("sunny" and "not sunny").
CTA	Polygonal edges	Variant of the LASMEA 2 algorithm alternating color and texture segmentation with an edge detection step.

Table 1: General description of the six algorithms involved in the assessment.

nor to the horizon. It provides a unique result that is well suited for error minimization.

The metrics have been computed on each image for every algorithm, which provides eight curbs (or more if we distinguish the right edge from the left edge) for each image sequence. Scalar measures can also be derived from these curbs, considering for instance the minimum, maximum and mean values as well as the variance over the whole sequence. Therefore, the amount of results is huge and we will only describe significant results that illustrate the analysis.

The examination of these various results enables us to outline general tendencies for each algorithm. Among the various metrics available, m_7 appears as a useful tool to get general indications about the behavior of an algorithm (indeed, global metric m_8 is often close to zero). For instance, metric m_7 clearly shows that some algorithms present frequent failures but are able to recover from difficult situations while others tend to provide very reliable results but sometimes suffer from irreparable failures. Metric m_4 is also interesting to evaluate the average precision of the algorithms and makes it possible to identify the images where the algorithms face difficulties. However, when an edge has not been detected by the algorithm, m_4 is arbitrarily set to the length of the image diagonal in order to penalize this result (this appends in Table 2).

Other metrics enable us to determine promising approaches among the algorithms. Some algorithms seem to be specifically tuned for one environmental condition while others tend to perform well on most sequences.

The various results also help us to underline complementary assets between the algorithms and to consider possible hybridations among the different techniques. For instance, one algorithm is a variant of a contour-oriented technique in which color and texture information has been added : the good results obtained by this new variant compared to the initial algorithm indicate that the fusion between contour-oriented and region-oriented techniques is probably a promising approach.

So far, the evaluation has been performed on four algorithms, namely the algorithms proposed by the LASMEA, the MINES and the CTA. Each algorithm has been tested on six general scenarios under three different weather conditions (sunny, cloudy and rainy). The metrics have been computed on each image for every algorithm, which provides 8 curbs (or more if we distinguish the right edge from the left edge) for each image sequence (see Fig. 3 for instance). Scalar measures can also be derived from these curbs, considering for instance the minimum, maximum and mean values as well as the variance over the whole sequence. Therefore, the amount of results is huge and we will only present a few significant results that illustrate the analysis (see Table 2 for instance).

The examination of these various results enables us to outline general tendencies for each algorithm:

- Among the selected approaches, the technique developed by the MINES globally yields the best results on scenario 4 (easy gravel road). Otherwise, it provides average results on most scenes. As shown in Fig. 3 for instance, although it presents frequent failures, it is able to recover from difficult situations: the metric periodically decreases to very low values be-

Algorithm	m_3 right	m_4 right	m_3 left	m_4 left	m_7	m_8
LASMEA 1	0.211 ± 0.017	32.46 ± 3.243	0.029 ± 0.005	61.97 ± 3.269	0.786 ± 0.010	0.006 ± 0.001
LASMEA 2	0.148 ± 0.008	86.95 ± 4.114	0.065 ± 0.005	52.04 ± 1.835	0.641 ± 0.008	0.008 ± 0.001
MINES	0.104 ± 0.006	84502 ± 7657	0.050 ± 0.004	87000 ± 7716	0.7143 ± 0.007	0.003 ± 0.0005
CTA	0.142 ± 0.009	38.80 ± 1.86	0.115 ± 0.0086	56.88 ± 2.4	0.78 ± 0.0059	0.017 ± 0.0021

Table 2: Measure m_3, m_4, m_7 and m_8 for sequence S5 cloudy

fore reaching an acceptable level again. This ability is probably induced by the failure test (based on a coincidence between prediction and estimation) which has been incorporated into the algorithm, thus activating a new initialization step whenever necessary.

- The first algorithm proposed by the LASMEA experiences trouble and seems to be the less reliable on most image sequences, except on the sunny sequence derived from scenario 4. Therefore, one can wonder whether the main parameter (the Mahalanobis threshold) was not specifically tuned for this kind of sequence. This behavior will be further investigated by studying the sensitivity of the algorithms to parameter tuning.

- The second algorithm developed by the LASMEA globally provides results which are similar to the algorithm proposed by the MINES. It experiences frequent failures as well (see Fig. 3), although these failures appear slightly less severe than those from the MINES. It is also capable of recovering from most failures. In this case too, it seems to be worth studying the effects of parameter variations.

- Finally, the approach proposed by the CTA tends to provide the best global results, although it often suffers from irreparable failures. These failures sometimes derive from a "segmentation fault" which may be due to an attempt to inverse a non-invertible matrix in the Kalman filter process. Another kind of failure simply leads to a progressive deviation of the road detection result away from the ground truth (see Fig. 3 for instance). Moreover, the results tend to show that a compromise has to be found between the rigidity (both in the temporal and geometric sense) and the stability of road model. Indeed, the CTA road template is sometimes too rigid to cope with abrupt bends and it cannot deal with "S"-shaped roads. However, when the road presents a classical shape, this rigidity seems to help it to ajust to the template. Similarly, the temporal stiffness of the road template (probably due to the Kalman filtering) prevents the algorithm from limiting the progressive deviation away from the real road edges.

The various results also help us to underline complementary assets between the algorithms and to consider possible hybridations among the different techniques. For instance, the CTA approach is basically a variant of the second LASMEA algorithm in which color and texture information has been added to contour-oriented techniques. Therefore, the good results obtained with the CTA algorithm compared to the original one tend to indicate that the fusion between contour-oriented and region-oriented techniques is a promising approach. Besides, it seems that the CTA algorithm could be improved using failure detection tests or recovering capabilities such as those developed by the MINES and the LASMEA.

5 EXPLOITATION OF THE RESULTS

As pointed out before, we have opted for a generic approach which makes it possible to evaluate very dissimilar algorithms relying both on various strategies and on very different road models. Moreover, the choice of the metrics also meets a genericity requirement so that the evaluation results can be used to select the best suited road detector whatever the final task or the command rules which enable the UGV to follow the road. However, even though the metrics help to reduce the initial data space (scalar measures on each image or on sequences with respect to reference road edges and algorithm results on each image), such a generic evaluation still provides numerous results to be exploited.

It should first be noted that the assessment results can be exploited at different levels, which enables us to adapt the amount of data to examine. Indeed, as shown in the previous section, the interpretation can be conducted at various levels:

- on different sets of images : the whole database, only general or special scenarios, easy or difficult scenes, a particular sequence or a given image, etc. ;

- using single or multiple metrics ;

- considering individual algorithms or comparing several ones.

It thus makes it possible to determine the strengths, weaknesses and complementarities between the algorithms at different levels.

Moreover, once the task is more precisely defined or if the command rules are given, we can select the most promising road tracking algorithms by choosing the most

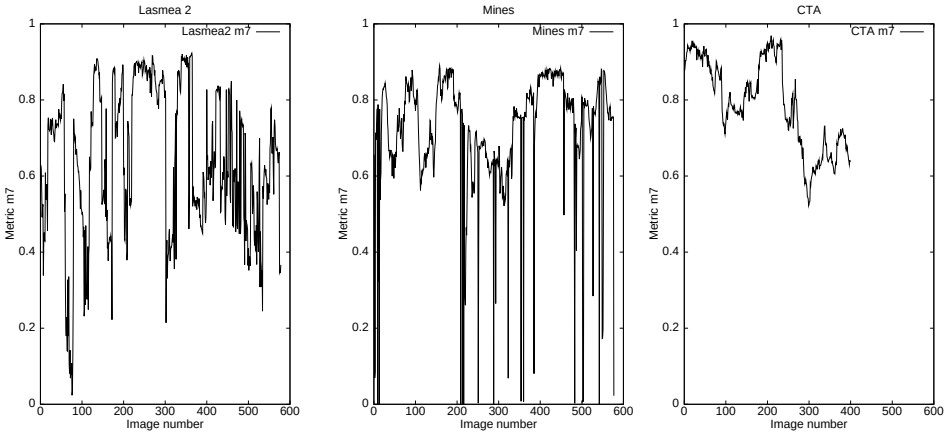

Figure 3: Measure m_7, sequence S5 cloudy for LASMEA 2 (left), MINES (center) and CTA (right).

relevant metric or combination of metrics for the task. We believe that the existing metrics already provide very useful information, so that in most cases, it would not be necessary to start the evaluation all over again, using metrics specifically designed for the task. To illustrate this idea, we propose the few following examples:

- If the aim of the system is simply to follow the road by staying on the central axis, we do not really care about the precision of the edge detection. What really matters is that the central road axis should be correctly recovered, so that the signed metric m_{4s} seems to be a good candidate
- If the system is dedicated to vehicle following while checking that it remains on a "navigable" area, the main concern is to detect the whole road surface. As a result, we tend to favor completeness, as well as correctness (to avoid the roadsides which are potentially dangerous).
- If the system aims at following a single road edge (for example if the vehicle needs to remain on the right lane to avoid vehicles coming from the opposite direction), the "buffer" metrics seem better suited than the surface metrics. As for the distance metrics, they may be too sensitive to small variations of the edge position (which cannot be detected correctly when the road models are too rigid).
- If the goal of the system is to detect the road edges with a good precision (for instance to determine how structured the road is, to analyze its quality or to determine the nature of the roadsides), the distance metrics seem to be the most relevant.

Furthermore, depending on whether the algorithms use region classification or edge detection, the use of corresponding surface-oriented or edge-oriented metrics may seem more natural.

Finally, we have seen that we can mimic human judgement by describing the algorithms as rather pessimistic or optimistic. We could go still further by fully automating the human judgement (to compare multiple variant of the same algorithm for example). In particular, this approach looks interesting in the case of a warning system that would raise on alarm if a pilot or a teleoperator tended to leave the "navigable" area. Indeed the most relevant quality metric for such a warning system might then be the human judgement itself. In the field of photo-interpretation, Letournel has defined a "qualitative objective metric" which meets such requirements. Indeed, this new kind of metric links numerical values (the objective quantitative metrics) with a semantic classification of the outcome of the algorithms (the human judgement) [10]. In her PhD thesis, this metric was dedicated to building extraction in aerial images but the metric definition could be adapted to road detection. The main steps then consist in :

- selecting significant images from the database and significant results of the algorithms on these images (good , bad and average results) ;
- collecting marks from human evaluators ;
- performing principal components analysis to reduce the initial metrics space (m_1, m_2, m_4, m_{4s}, m_5 and m_6 for instance) ;
- building performance measures that reflect the human judgement. The first method relies on statistical analysis (canonical analysis) to select an optimal subset of features corresponding to the human notation and builds a similarity measure which is a linear combination of metrics. The second method relies on fuzzy logic to select a new subset of metrics and build a new similarity measure.

- Both similarity measures can be combined to propose a final performance measure that discriminates between good and bad segmentations.

It should be noted that these qualitative objective metrics can be computed using our initial "quantitative" metrics results, so that no new quantitative comparison between the reference and result edges is needed.

6 LESSONS LEARNED ABOUT THE ASSESSMENT METHODOLOGY

6.1 *Constitution of the database*

First, this evaluation allowed us to gain insight about the constitution of the database:

- It turned out extremely difficult to find areas on the proving ground that could satisfy simultaneously all the requirements for the scenarios. In particular, most roads presented numerous intersections and other difficulties, which was to be avoided for the "easy" scenarios. Therefore, it may be useful to plan the image acquisition on a very large area containing numerous and various kinds of roads, or to share out this acquisition among different participants.

- As described above, the CTA has developed a dedicated interface to facilitate the ground truth marking by the human expert. Although this interface seems very useful, several possible improvements have been identified. For instance, it seems interesting to display simultaneously the current and the previous image, so that the operator can define vertices for the polygonal lines that always correspond to the same points in the scene. Moreover, to accelerate the ground truth definition process, it would be helpful to use a reference algorithm which could be manually controlled and stopped (as currently investigated by the NIST).

- Finally, it seems interesting to split the database into two or three parts: a learning database and a development database with ground truth could be used by the laboratories to train and test their algorithms, whereas an evaluation database would be dedicated to the final assessment. In the scope of our study, such an approach was not required since the algorithms did not need an extensive training, but it should be taken into account for further extensions.

6.2 *Lessons learned about the choice of the metrics*

We also learned a few lessons about the metrics:

- In section 4, we have already described the assets and limitations of some metrics. More generally:
  - The "buffer" metrics m_1 and m_2 tell how valid the detected road edges are, whatever the slight variations in the precision. However, they do not indicate the directions of errors.
  - The surface metrics m_5 and m_6 outline general tendencies for the algorithms (they seem more stable than the edge metrics [3]) and may provide global indications about the direction of errors in some cases. However, they are sensitive to slight variations of the edge localization and cannot distinguish between right and left edges.
  - The unsigned distance metric m_4 points out the images where the algorithm fails and provides information about the precision. However, it does not indicate the type of error (false positive / false negative or pessimistic / optimistic).
  - As for the signed metric m_{4s}, it can determine the offset for the central axis of the road and combined to m_4, it may show the direction of the offset on each road edge. However, since positive and negative errors compensate for each other, the mean offset value is not significant with the signed metric.

- Some metrics had to be slightly modified in order to cope with the large variety of algorithms. In particular, some techniques consider a fixed horizon, which induced a systematic error on the completeness metrics within the upper part of the road. As a result, we introduced a weighted sum into the distance metrics (m_4 and m_{4s}) so as to favor the lower part of the image and enable a more consistent comparison between the different algorithms.

- It seems interesting to define temporal metrics that would take into account the temporal consistence of the road detection between two consecutive images (so that the offset between the reference and the result remains on the same side, for instance). Global temporal metrics could be computed using the existing static ones (m_{4s} for instance) but more precise metrics might require new computations.

- Some metrics depend on parameters: for instance, the width of the tolerance buffer can be modified. In the current evaluation, it was set quite arbitrarily to 12 pixels, taking into account the precision of the ground truth (5 pixels) and adding a few pixels to allow slight variations for the algorithm. However, this width could be set more rigorously, studying the impact of an error in the image processing results on the control algorithms for the UGV for instance (which implies that to remain generic, the evaluation needs to be performed for different values of this width).

- It would be interesting to study the metric behaviors more thoroughly. At first, we could globally examine

their variations (compute the histograms and variation coefficients over the database). We could also analyze their correlation in an attempt to reduce the metric space [10]. Finally, we could study analytically the sensitivity of these metrics. However, this still looks quite difficult to formalize.

6.3 *Lessons learned about the whole methodology*

Since we decided to perform the evaluation using binaries provided by the laboratories, it sometimes turned out difficult to solve small practical problems when running the algorithms, which appeared as mysterious black boxes. An alternate solution would consist in sending the database images (without the ground truth) to the laboratories and let them compute the results themselves, taking advantage of potential dedicated hardware. However, this would allow the participants to tune their algorithms on the image sequences and might introduce biases on the assessment (even though in some cases, these biases could eventually be detected). To tackle this problem, one could imagine sending the images through internet and allowing only a very short time for the laboratories to send their results back. However, if this approach is widespread in the speech processing community, it seems more difficult to apply to image sequences which represent huge amounts of data and might saturate the networks.

7 CONCLUSION

This article describes the results of a quantitative performance evaluation concerning road detection and tracking algorithms. It also presents the lessons we learned about the assessment methodology, especially about the database constitution and the metrics definition. It is true that this kind of evaluation, which only takes into account the image processing task, cannot replace a global system evaluation : in particular, for an automatic road following task, other elements from the intelligent vehicle need to be assessed, such as the command and control strategy. However, the methodology we propose can also take into account external elements such as the use of other kinds and sensors, the use of 3-D vehicle models that take advantage of proprioceptive information, etc. Indeed, the contribution of these external elements to the image processing task can be measured quantitatively, these external data or models being often used to guide the vision algorithms.

To conclude, our first results applying the methodology look very encouraging. Despite the slight adaptations to the metrics, we succeeded in understanding the general behaviors of the road detectors. Therefore, after this first experience, we believe that it would be worth making this kind of evaluation more systematic. Indeed, in the development phase of an algorithm, such tests make it possible to orient the design choices, for instance by comparing pairs of approaches that would only differ by a single element (the geometric road template for example). The definition of modular algorithms would also help testing variants in this prototyping process. In the evaluation phase, such a methodology enables to test quite rigorously the contribution of the new algorithm with respect to existing approaches. Moreover, a systematic evaluation would help standardizing the input and output formats for the algorithms, which would make the evaluation process easier.

References

[1] A. Broggi, M. Bertozzi, A. Fascioli, and G. Conte. *Automatic Vehicle Guidance: the experience of the* ARGO *autonomous vehicle.* World Scientific Publishing, 1999.

[2] D. DeMenthon. Inverse perspective of a road by local image matches and global 3D optimization. Technical Report CAR-TR-210, Univ. of Maryland, 1986.

[3] D. Dufourd and A. Dalgalarrondo. Performance evaluation of road detection and tracking algorithms. Workshop PerMIS'O2, Gaithersburg (MD), august 2002.

[4] D. Dufourd and A. Dalgalarrondo. Quantitative evaluation of image processing algorithms for ill-structured road detection and tracking. SPIE AeroSense'O3, Unmanned Ground Vehicle Technology V, Orlando (FL), april 2003.

[5] J. Everson, E. Kopala, L. Lazofson, H. Choe, and D. Pomerleau. Sensor performance and weather effects modeling for intelligent transportation systems applications. In *Intelligent Vehicle Highway Systems (Proc. SPIE vol. 2344)*, pages 118–128, 1994.

[6] A. Guiducci. Parametric model of the perspective projection of a road with applications to lane keeping and 3D road reconstruction. *Computer Vision and Image Understanding*, 73(3), Mar. 1999.

[7] M. Herman, S. Szabo, K. Murphy, D. Coombs, T.-H. Hong, H. Scott, N. Dagalkis, K. Goodwin, and J. Albus. Recomandations for performance evaluation of unmanned ground vehicle technologies. Technical Report NISTIR-5244, NIST, Aug. 1993.

[8] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(7):673–689, 1996.

[9] K. C. Kluge. Performance evaluation of vision-based lane sensing: some preliminary tools, metrics and results. In *IEEE Conf. on Intelligent Transportation Systems*, 1997.

[10] V. Letournel. *Contribution à l'évaluation d'algorithmes de traitement d'images.* PhD thesis, ENST Paris, CTA / Centre Technique d'Arcueil, octobre 2002.